RGB-D Sensing based Human Action and Interaction Analysis: A Survey

Bangli Liu, Haibin Cai, Zhaojie Ju, and Honghai Liu*

Intelligent Systems and Biomedical Robotics Group, School of Computing, University of Portsmouth, UK

Abstract

Human activity recognition has been actively studied in the last three decades. Compared to human action performed by a single person, human interaction is more complex due to the involvement of more subjects and the interdependence between them. Recently, motivated by the remarkable success of deep learning techniques, many learning-based feature representations have been developed for activity recognition. This paper provides a comprehensive review of human action and interaction recognition methods, covering both hand-crafted features and learning-based features, with a special focus on data captured by RGB-D sensors. Furthermore, this review reveals practical challenges in human activity analysis along with their promising solutions and potential future directions.

Keywords: Survey, RGB-D sensing, Action recognition, Interaction recognition

1. Introduction

Human activity recognition has attracted increasing attentions due to its wide applications in video surveillance, elderly care, virtual reality, and human-machine interaction. According to the complexity of human activities, they can be broadly classified into the following four categories [1, 2]: atomic action, human-object interaction, human interaction, and group activity. This paper mainly focuses on atomic action performed by a single person, and interaction between human and human.

*Corresponding author.
Email address: honghai.liu@port.ac.uk

The release of cost-effective RGB-D sensors has motivated plenty of RGB-D data based human activity recognition methods being proposed. Table 1 lists some specifications of the most popular RGB-D sensors. As a pioneer, the Kinect v1 sensor, which can jointly output color, depth, and skeleton data at 30fps, has been actively explored in many areas such as human activity recognition, facial behavior analysis, and 3D reconstruction. However, its limitations such as a maximum sensing range of 4 meters and unsuitable for outdoor scenarios are also well identified. Recently, the newly developed Intel Realsense sensors have overcome the outdoor limitation and can sense a longer distance at 10 meters, paving the way for its broad application.

Table 1: Properties of RGB-D sensors.

Name	RGB	Depth	Scene	Range	Year
Kinect v1	640x480(30 fps)	640x480(30 fps)	Indoor	0.4-4 meters	2011
Kinect v2	1920x1080(30 fps)	512x424(30 fps)	Indoor	0.4-4.5 meters	2013
Xtion PRO	1280x1024(30 fps)	640x480(30 fps)	Indoor	0.8-3.5 meters	2012
Xtion 2	2592x1944(15 fps) 1920x1080(30 fps)	640x480(30 fps)	Indoor	0.8-3.5 meters	2017
Intel RealSense D415/D435	1920x1080(30 fps)	1280x720(90 fps)	Indoor Outdoor	0.16-10 meters/ 0.11-10 meters	2018

RGB-D sensor-based human activity recognition is a fundamental technique for many practical applications. For example, in healthcare scenarios, it could facilitate the monitoring and analysis of patients' motion rehabilitation process by releasing the requirement of wearing sensors. Similarly, by recognizing elderly people's emergency, such as falling down, it can provide the necessary information to inform an assisted robot or corresponding organizations [3]. Regarding education scenarios, this technology could be used to improve the autonomy of the robots, thus enables them to teach children with autism spectrum disorder social interaction skills [**?**]. In sports fields, human activity recognition can be used to record and analyze the performance of athletes, which is beneficial for their further improvement. In human-robot interaction or collaboration scenarios, robots could perform desirable activities by interpreting human intentions. Human activity recognition could also be used in virtual reality related applications, which allows users to have natural interactions with an augmented environment.

The main challenges of human activity recognition are online adaption, occlusion, viewpoint variations, execution rate variations, and biometric changes. Online adaption is an ability to detect the occurrence of actions and provide an instant classification in continuous video streams, which is also referred as online activity recognition. Compared to traditional action recognition which typically focuses on classifying the manually trimmed actions and giving the result after the event, online action recognition is more challenging in that the occurrence of actions needs to be automatically detected and the recognition needs to be conducted in situations where only partial actions can be observed. The second challenge is occlusion, where inter-occlusion and self-occlusion might cause difficulties in the detection of different body parts [4]. Viewpoint variations and biometric changes caused by different human body size, appearance, shape, and distance from the sensor to subjects will lead to large intra-class variability and affect the performance of algorithms. The execution rate variations may also occur due to different performing styles and speeds.

Several survey papers have summarized the research on human activity recognition using RGB-D sensors [5–11]. Zhang *et al.* [5] provided an overview of existing RGB-D action datasets. Chen *et al.* [8] reviewed depth-based human action recognition approaches. Lu *et al.* [12] presented a review for Kinect sensor based motion recognition applications. Skeleton-based action recognition methods with different anatomy are reviewed in [6] and [10]. There are also several reviews of activity recognition for both skeleton and depth images such as [9] and [7]. However, at the time of writing, there is no survey specifically focused on RGB-D based human interaction recognition, which is popular in daily life and has received increasing attention. To fill this gap, this paper presents a comprehensive overview of RGB-D sensing based human action and interaction recognition, covering both hand-crafted methods and learning-based methods. Although Zhu *et al.* [11] also reviewed both types of methods, they focused on RGB data-based human activity recognition. Moreover, this paper presents a discussion for practical challenges in human activity recognition and their promising solutions in order to inspire future research.

The main contributions of this survey are summarized as follows:

1) A thorough overview of human action and interaction recognition using RGB-D

sensors is presented.

2) A comprehensive analysis of both hand-crafted and deep learning based methods is conducted.

3) The challenges of human activity recognition using RGB-D data and existing solutions are discussed.

Fig. 1 shows the structure of this paper which is organized as follows: Section 2 reviews the hand-crafted human action recognition algorithms. Hand-crafted features based human interaction recognition methods are introduced in Section 3. Section 4 reviews deep learning based human activity recognition methods. Section 5 demonstrates the challenges and relative solutions for human activity recognition. Section 6 provides a comparison between hand-crafted and deep leaning-based representations along with a discussion of their performance on the most commonly used datasets. Finally, Section 7 concludes this paper and discusses the future directions.

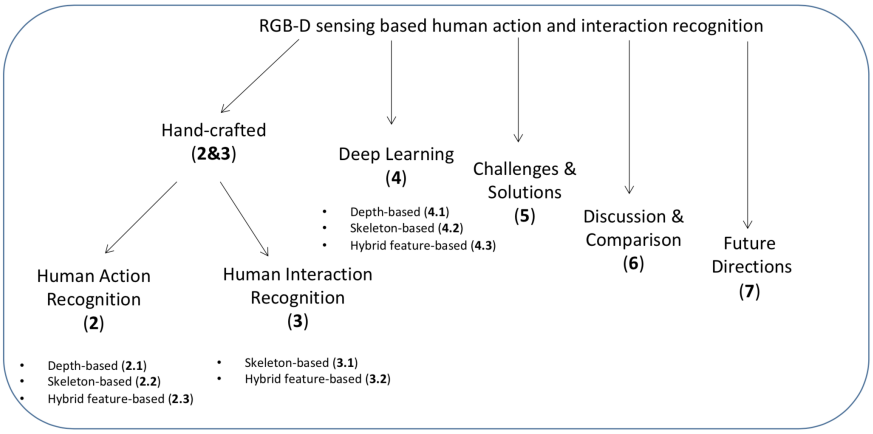

Fig. 1: Structure of this paper.

2. Hand-crafted Features based Human Action Recognition

The existing RGB-D based human action recognition methods can be classified into three categories depending on the used data modality: depth-based methods, skeleton-based methods, and hybrid feature-based methods.

2.1. Depth-based Methods

The depth images, which store the Euclidean distance between the sensor and points in the scene, make it easy to extract human bodies from the cluttered background. Some researchers [13–17] proposed to project the 3D depth information onto three 2D orthogonal planes corresponding to the front, side, and top view for feature extraction. Li *et al.* [14] extracted 3D representative points of the body silhouette from these planes to model postures for recognition. However, as pointed out in [18], dealing with the large amount of extracted 3D points requires significant time and memory consumption. In [18], Depth Motion Maps, i.e., DMMs, were generated by stacking depth maps with a threshold between two consecutive frames and then HOGs were computed from these DMMs to characterize human motions over the whole sequence. Later, Chen *et al.* [13] argued that replacing the final HOGs representation with a concatenation of DMMs can not only reduce the computational cost but also achieve better recognition performance. To address the speed variations in actions, a multi-temporal DMM representation [16] was proposed to extract the shape and motion cues from different lengths of depth segments. The temporal information among frames was also restored in this representation by introducing a weighting function into depth sequences. Bulbul *et al.* [15] improved DMMs by implementing the contourlet transform with a multi-scale and multi-directional analysis to enhance the shape characteristic of DMMs. One limitation of these methods is that they do not take the neighborhood of 3D points into account and thus might discard useful information.

The surface normal vectors calculated using a group of 3D points can be used to describe the shape and motion information [19–21]. Oreifej *et al.* [19] proposed to divide the depth sequences into many spatiotemporal cells to compute the Histogram of Oriented 4D, i.e., HON4D, which depicts the distribution of the surface normal orientation, for action recognition. In [22], HON4D extracted from each action video was used to build tensor representations in a tensor subspace [23] to preserve discriminant and local information. Similarly, super normal vector [20], i.e., SNV, was calculated by grouping local hypersurface normals to create the low-level polynormal, which further preserves the correlation among local normals in the polynormal and achieved a better recognition rate. Slama *et al.* [21] modeled the normal vector orientation sequence fea-

ture as subspaces lying on Grassmannian manifold and employed a probability density function for classification.

Alternatively, some researchers proposed to segment the depth data to interest areas, from which compact features were extracted for action recognition. For example, Wang *et al.* [24] constructed random occupancy patterns feature from 4D subvolumes randomly sampled in depth map sequences to gain the robustness towards occulsions. Xia *et al.* [25] utilized the depth cuboid similarity to depict the local feature around the spatio-temporal interest points extracted from depth videos. In [26], the spatial relationship among selected joints with discriminative shape and movement was used to build the depth context descriptor for final action recognition. Liu *et al.* [27] generated motion-based and shape-based spatial-temporal interest points (STIPs) using the motion and shape information from depth data respectively. Then, a two-layer bag-of-visual-words model was introduced to describe the local appearances and the distribution of STIPs. One limitation of these approaches is that detecting interest regions through the whole depth sequence requires extra computational cost.

2.2. Skeleton-based Methods

The release of RGB-D sensors such as Kinect and Xtion enables us to obtain 3D positions of body joints from depth images [28] in real time performance, encouraging many skeleton-based methods being proposed. They can be further divided into trajectory-based and pose-based algorithms.

2.2.1. Trajectory-based Algorithms

Trajectory-based algorithms explore characteristics of the spatiotemporal trajectory of skeleton joints to identify a set of distinctive features [29–33]. Gowayyed *et al.* [29] proposed a 3D trajectory descriptor, which concatenates three 2D projections of the whole skeleton sequences, to represent the movement of each joint. In [30], actions were modeled by computing the similarity and dynamics information of joint angles. Qiao *et al.* [32] applied a trajectorylet based on local feature representations, which constrained the dynamic characteristic of actions from the entire sequence to a short temporal range, to capture ample static and dynamic information of actions. Devanne

et al. [33] modeled motion trajectories of actions as points in the open curve shape space by transferring the 3-D coordinates of skeleton joints to a Riemannian manifold. Then, the action classification was achieved by computing the similarity between the shape of trajectories in the manifold. Guo *et al.* [34] decomposed the human body skeleton into five parts and proposed a gradient variation based feature to represent the 6D rigid body motion trajectories. After coding the skeleton representations into a sparse histogram, a SVM with chi-square kernel was used for action recognition.

2.2.2. Pose-based Algorithms

Compared to the trajectory-based approaches, pose-based approaches focus more on key poses characterized by the skeleton point distribution or its surrounding body parts. Features such as joint locations, joint angles, 3D geometric relationships between body parts are often directly employed as advantageous representations of activities [35–38]. In [35], a histogram of 3D joint locations in a spherical coordinate system was proposed to describe key human postures. Then, a discrete hidden markov model was utilized to explain the temporal evolution. Pazhoumand *et al.* [36] used joint angles and the relative motions between joints to depict body poses and the relationships between joints in the time domain. Instead of using the movement of all skeleton joints, Eweiwi *et al.* [39] focused on mining discriminative joints with apparent motion property. Several discriminative joints were determined by partial least squares, whose location information, velocity and the movement normals were encoded as poses during a short video period. Chaaraoui *et al.* [40] used a matching between action sequences by Dynamic Time Warping, i.e., DTW, for action recognition, where an evolutionary algorithm was proposed to select the optimal set of skeleton joints to form sequences of key poses for each action. Vemulapalli *et al.* [41] made use of the rotations and translations among five body parts to model their relative 3D geometry relation, with which human motion was encoded as curves in the Lie group. This method can reveal the concurrence of body parts, whereas the isolation of body parts may be difficult when there is overlapped areas among body parts.

2.3. Hybrid Feature-based Methods

The association of multi-modal data such as skeleton data, color, and depth images might improve the recognition performance. Many hybrid features tend to extract the corresponding depth information around skeleton joints [42–45], or combine the features from joints and depth images directly [46–50]. Wang *et al.* [42, 44] proposed the local occupancy pattern (LOP) feature to describe the appearance around each joint by recording the depth information in its neighborhood. Ji *et al.* [50] partitioned the human body to several motion parts by embedding the skeleton data into depth sequences. Local features extracted from these motion parts were aggregated into a discriminative descriptor. The depth information of objects around joints was also associated in [43] as the low-level layer of a hierarchical HMM. Zhu *et al.* [48] coupled the motion depending on points of interest and spatial information using a random forests-based fusion strategy. Yang *et al.* [46] proposed a depth map based accumulated motion energy function to select the discriminative skeleton frames to remove noisy frames and reduce computational cost. After the calculation of eigenjoints, they used non-parametric Naïve-Bayes-Nearest-Neighbor to classify multiple actions.

Apart from the combination of skeleton joints and depth images, some researchers also consider RGB information [51–56]. Sung *et al.* [51] employed skeleton joints to model motion features and extracted HOG features from regions of interest in both RGB and depth images to characterize the appearance cues. A coupled hidden conditional random fields model [52] was proposed to learn the latent correlation between visual features from both RGB and depth source. In this model, the temporal conext within individual modality is preserved while learning the correlation between two modalities. Kong *et al.* [53, 55] projected features from RGB and depth images into a shaped space and independent private spaces for action recognition, which indicates that knowledge and correlation from different sources could be shared with each other to reduce noise and improve the action recognition performance.

3. Hand-crafted Features based Human Interaction Recognition

The majority of existing RGB-D data based human interaction recognition use features from skeleton data or combine features from different channels, while few approaches are only based on depth images. Therefore, we classify them into two categories: skeleton-based methods and hybrid feature-based methods.

3.1. Skeleton-based Methods

Many human interaction recognition algorithms extract features from each individual's joints and their interactive joints to represent the motion relation over time. Yun *et al.* [57] used the joint distance and movement between all pairs of joints of two-person, the geometric relationship between joints and planes, and velocity features to represent the motion. Then, a Multiple Instance Learning classifier was proposed to handle irrelevant frames in the trained data. In [58], interactions were disjointed into topics and a hierarchical model was employed to exhibit the correlation among low-level features, topics, and activities. Mining the essential interactive pairs helps to remove redundant information from the inactive body parts and improve the computational performance. For example, Ji *et al.* [59, 60] applied the contrast mining method to extract the most active body part pairs for each interaction class. Wu *et al.* [61] proposed a human interaction feature descriptor by utilizing the static, dynamic, and direction properties of the skeleton data. They addressed the interaction recognition problem by using a Sparse group Lasso penalty enhanced linear Model (SLM).

Some scholars transformed the interaction problem to the single person action recognition problem [62, 63]. The interaction between players was decomposed into two single individual actions in a computer gaming environment in [62], where each player's action was trained and classified separately. Hu *et al.* [63] firstly identified the most active person according to the following two rules: the person acts firstly or the person with greater motion at the beginning short frames.Then, the action of the active person was used for human interaction recognition. Unlike the methods mentioned above, Coppola *et al.* [64] utilized features from two individuals and the relationship between each other for different purposes. They treated physical proximity features

learned from social interaction as prior knowledge and built a multivariate Gaussian distribution to estimate the distribution of each interaction category.

3.2. Hybrid Feature-based Methods

Features from different modalities can provide extra information for activity recognition. Gori *et al.* [65] built a bounding box around the human body to remove most of the redundant information of the different modalities. Then, a matrix called relation history image was proposed to depict the local relations, which contains Euclidean distances of joint pairs and comparison of depth value between pixels. Similarly, Van *et al.* [66] explored shape and movement features for each interactive person from bounding boxes where the interaction happens, and they merged the information of joints with poselets to select key frames in the training stage. Xia *et al.* [67] combined skeleton joints-based postures, motion information described by 3D optical flow, and local appearance feature around spatiotemporal interesting points in both RGB and depth data for interaction recognition. Alazrai *et al.* [68] used the motion direction and distance between two persons to describe the relationship of body-parts and further extracted local shape information from the bounding box around body parts. The final feature descriptor was formed by concatenating all these features. Trabelsi *et al.* [69] proposed to jointly use the distance property of the 3D skeleton and the dense optical feature extracted from the color and depth images for interaction recognition.

4. Deep Learning based Human Activity Recognition

Unlike the hand-crafted methods where specific types of features need to be designed to distinguish human action and interaction recognition, most of the deep learning based methods code human action or interaction information directly into a map and then resize the map to a fixed size or directly concatenate the representation of each person as an input of networks for recognition. Therefore, to outline the key difference between different deep learning based methods, this paper doesn't specifically separate existing deep learning based human activity recognition methods into single human action and human interaction at this stage. Following the same taxonomy with

the hand-crafted methods, the research reviewed in this section can be grouped into three categories: skeleton-based, depth-based, and hybrid-feature based.

4.1. Skeleton-based Methods

The skeleton-based methods can be further separated into CNN based methods and RNN based methods according to the adopted deep learning structure.

4.1.1. CNN

Most of the CNN based action recognition methods focus on transforming the positions or trajectories of skeleton joints into images and then adapting CNN for classification. In [70], a linear interpolation function is utilized to construct four joint distance maps from the 3D distance information and three orthogonal 2D planes projected by 3D skeleton joints. The action was classified by using the constructed distance maps together with AlexNet. Ke *et al.* [71] constructed three clips of gray images using the relative positions between the skeleton joints and four manually defined reference joints. By feeding the gray images into a pre-trained VGGNet and developing a multi-task learning network, the spatial structural information was incorporated for action recognition. Observing that the image resizing operation might introduce extra noise for the network, Liu *et al.* [72] proposed to directly input a skeleton image to a modified Inception-ResNet CNN architecture for action recognition. The drawback of this method is that the assumption of each action has a fixed number of frames as input. The spatiotemporal information of 3D skeleton sequences was encoded into three joint trajectory maps according to three different views (i.e., front, top, and side) in [73, 74]. The action was classified via a late fusion of three ConvNet trained from trajectory maps. To ease the variations of skeleton sequences in the spatial and temporal domain, Xie *et al.* [75] recalibrated action sequences temporally in a residual learning module and then modeled their spatial and temporal information using CNNs for final action recognition.

Different with previous methods, Yan *et al.* [76] proposed a multi-layer graph neural networks, where the graph nodes consist of joint coordinates and estimation confidences, to automatically learn the spatial-temporal pattern of the skeleton data.

Huang *et al.* [77] employed a neural network architecture to learn a temporally aligned Lie group representations [78] for action recognition, which demonstrated that the non-Euclidean Lie group structure can also be incorporated by the deep learning structure. Observing that skeleton joints might be unreliable due to the occlusions and noisy backgrounds, Liu *et al.* [79] proposed to concatenate the 2D coordinates to a pose estimation map frame by frame, from which a body shape evolution image and a body pose evolution image were constructed to interpret action segments.

4.1.2. RNN

Compared to CNN, RNN could effectively model the temporal information. Most of the existing RNN based methods employ Long Short-Term Memory (LSTM), which solves the gradient vanishing problem by utilizing a gating mechanism to determine the memory length of the input sequence, to process long action sequences. Thus, instead of converting motion information to images, RNN based methods tend to directly use joints or the connection of joints as input.

Veeriah *et al.* [80] proposed a differential RNN by adding a gating into LSTM to model the dynamics of salient motions. Various hand-crafted features concatenated from successive frames were fed to the proposed LSTM structure. Du *et al.* [81, 82] proposed an end-to-end hierarchical RNN which fuses the feature extracted from five human body parts for action recognition. However, as pointed out in [83], the relationship between non-adjacent parts was ignored in this method. Shahroudy *et al.* [84] utilized the human body structure to build a part-aware LSTM. By concatenating part-based memory cells, the non-adjacent parts relations were learned from the 3d skeleton sequence. Mahasseni *et al.* [85] employed the regularized LSTM on top of a deep convolutional neural network for RGB video based action recognition. Assuming extra 3D skeleton data can complement the lost information in the video, they proposed to regularize the network by using the 3D skeleton sequence from a few actions. Zhu *et al.* [86] fed the skeleton joints to a deep LSTM network with mixed-norm regularization term to learn co-occurrence features for action recognition. They further applied an internal dropout method to the LSTM neurons in the last LSTM layer to learn complex motion dynamics. Zhang *et al.* [83] explored various geometric relational features

among all joints and used a stacked three layers LSTM for action recognition. Observing the lost information in the transforming of 3D skeleton joints to the person-centric coordinate system, Zhang *et al.* [87] proposed a view adaptive RNN with LSTM structure to deal with the viewpoint variations. Liu *et al.* [88] developed a global context-aware attention LSTM to selectively pay attention to informative joints in each frame with the help of the global memory cell. The attention ability was further improved by using a recurrent attention mechanism, which improved the recognition performance by reducing the noise of the irrelevant joints.

Unlike the previous RNN based methods where only the temporal domain of the skeletons are modeled, Liu *et al.* [89] proposed a tree-structure based traversal method to handle the spatial adjacency graph of the body joints. A trust gate was also proposed to remove noisy joints and deal with the occlusion in the 3D skeleton data. Similarly, Song *et al.* [90] proposed to add joint-selection gates in the spatial attention model and frame-selection gates in the temporal model for action recognition. Wang *et al.* [91] proposed a two-stream RNN architecture which jointly models the spatial articulated property and the temporal dynamic of skeletons. The additional spatial RNN modeled the spatial dependency of joints by considering human body kinematics. Si *et al.* [92] represented each body part as nodes in a residual graph neural network to capture the structural relationship between body parts at each frame. Then, a temporal stack learning network with three skip-clip LSTMs was introduced to model the temporal evolution of joint sequences.

4.2. Depth-based Methods

Depth image sequences might not be suitable to be the direct input of the most existing CNN models which are specifically designed for color images. Therefore, Some researchers proposed to extract hand-crafted features from depth sequences by stacking shape and motion features over the whole video and then convert them to texture images by encoding depth information. The generated texture images enable the use of existing models pre-trained on large scale image recognition or segmentation datasets with the finetuning operation to achieve satisfactory results. Wang *et al.* [73] encoded the DMMs feature [18] into Pseudo-RGB images by converting the spatial and

temporal movement information into textures and edges. Three independent ConvNets corresponding to three viewpoints were trained and the final recognition result was assigned by fusing the three generated class scores. Rahmani *et al.* [93] proposed to learn a view-invariant human pose model from depth sequences. Each frame of real depth videos was input to the CNN model to learn a view-invariant and high-level feature space, and then new human poses captured from unknown views were transferred to this space to achieve a cross-view action recognition.

4.3. Hybrid Feature-based Methods

Some researchers proposed to learn multi-modal features via separate networks for action recognition [94–98]. Zhang *et al.* [95] proposed to use 3D convolutional neural networks (3DCNN) [99, 100] and bidirectional convolutional long-short-term memory networks to learn spatial-temporal information from multi-modal data. The final gesture recognition was achieved by throwing the jointed multi-modal features to a linear SVM classifier. Kamel *et al.* [101] proposed to encode the consecutive depth maps and skeleton points into two separate images and further used three different combination settings to train three separate CNNs for action recognition. Wu *et al.* [96] developed a Deep Dynamic Neural Networks (DDNN) for gesture recognition with multi-modal inputs. The DDNN includes a Gaussian-Bernouilli Deep Belief Network to explore dynamic features from skeleton sequences, and a 3DCNN to extract spatial-temporal features from RGB and depth images. Instead of fusing results from each separate ConvNets, Wang *et al.* [102] proposed scene flow to action map to combine features from RGB and depth channels as the input to ConvNets. In [103], a privileged information-based RNN framework was investigated for action recognition by using depth sequences and skeleton joints. Liu *et al.* [104] proposed to learn high-level features from raw depth images and low-level features such as the position and angle information from skeleton joints. The two types of features were fused and inputted to SVM for action recognition.

5. Challenges

Human activity recognition involves addressing many challenges such as viewpoint and biometric variation, occlusion, various execution rates, and online adaption. This section will describe the challenges and review the efforts done to address these challenges.

5.1. Viewpoint Variation and Biometric Variation

The appearance of an action might change dramatically in different viewing angels and positions. The 3D skeleton data has an intrinsic property against the change of viewpoints [35–37, 41, 42, 105]. Most of the skeleton-based methods transform 3D joint coordinates from the world coordinate to a person-centric coordinate to achieve view-independent action recognition [35, 41]. An orientation alignment strategy was used to eliminate the influence of human body orientation by rotating joints plane to a certain plane [42]. On the other hand, most of the depth-based methods suffer from the dramatic shape and appearance change in different views. To learn view-invariant features through CNN models for captured depth maps, multi-view data is synthesized by rotating virtual cameras around the subject [73] or augmented by synthetically fitting 3D human models to real motion data and then producing poses from different viewpoints[93]. Similarly, Wang *et al.* [106] rotated the depth data in 3D point clouds in different angles to deal with viewpoint invariance.

Biometric variation is caused by many factors such as various body size, distance of the sensor related to the object, etc., which can result in different body shape or appearance. This might affect the performance of feature descriptors, especially for those based on shape or appearance characteristics. Various body size was typically tackled by normalizing the human body with one particular part of each human [31, 45, 107].

5.2. Occlusion

The cluttered surrounding or overlapped body parts might result in occlusions which make it a great challenge for human action recognition. This phenomenon becomes more serious when it comes to human interaction, where people can be occluded

by each other and oneself. The occlusion in human interaction also makes it difficult to isolate individuals and extract features from each unique person.

Most methods estimated the invisible parts according to previous frames information or the visible parts. Hsieh *et al.* [108] separated the occluded body parts by particle filter and triangulated them to triangular meshes, which then were re-labeled to repair the incomplete shape using a template re-projection technique. Probabilistic graphical model in a markov random field was utilized to measure the occluded state of body parts under self-occlusion in [109]. To address the frequent occlusion and feature-to-object mismatching occurring in close human interaction, Kong *et al.* [4, 110] proposed a patch-aware model, where supporting regions of each interacting subject were learned at patch level.

5.3. Action Duration Variation

The different action duration caused by various performing speed and habit of subjects might result in different dimensions of features, which cannot be the direct input of typical classifiers, such as SVM and kNN. A common solution is to use interpolation operation to unify the length of activity videos. Apart from this, DTW [41] and temporal pyramid models [29, 42] were popularly applied to make sure the same length of each sequence. The probabilistic graphical models such as HMM [43, 63], Bayesian networks, conditional random fields [111], and hidden conditional random fields [112], can be used to represent actions by the probability relation between states.

In CNN-based methods, a single color image is produced by encoding the depth or skeleton sequence frame by frame and further resized to a fixed size [70, 71]. Although this image resizing operation can tackle the temporal duration problem, it might introduce extra noise for the network. On the other hand, RNN or its variants can also be used to effectively model data sequences by exploring the temporal dependencies among frames [80–82].

5.4. Online Activity Recognition

Online activity recognition is quite challenging in that action detection and recognition need to be conducted simultaneously with a low latency so that the system can

provide an instant response. For example, the assisted robot should be able to provide immediate help for the elderly people if they are going to fall down.

To localize the action, most of the early works use a probability/energy-based threshold to detect the boundary or key poses of each action. For example, Zhu *et al.* [113] identified transit motion features between two continuous poses in training phase, and the online classification was achieved by comparing likelihood probabilities in the MEMM model. There are some methods executing segmentation according to the clip-level or frame-level labeling approach [114–116]. Wu *et al.* [115] clustered daily life clips to several action-words, with which an action-topics model was learned to reflect the co-occurrence and temporal relations. The action segmentation was realized according to the change of action topics between consecutive clips. Sliding window is also a popular and compact technique for online action recognition [61, 117], by which a video stream is usually divided into a set of overlapped segments and then classification is conducted in each segment.

Apart from the classic sliding window strategy, some deep learning based methods address this problem by developing different architectures. Molchanov *et al.* [118] proposed a recurrent 3DCNN to simultaneously perform classification and localization of hand gestures from continuous depth, color, and stereo-IR data sequences. Shou *et al.* [119] present to address action temporal localization via multi-stage CNNs, which includes identifying candidate segments that may contain actions, action recognition, and temporal boundary localization. Recently, RNN and its variants (e.g., LSTM) have been drawing attention for online action recognition [120, 121], owing to its appealing capacity of modeling temporal dynamics of sequences.

6. Discussion

This section provides a discussion for both hand-crafted methods and deep learning methods in terms of adopted classifiers, accuracies, and solutions to each challenge. Seven commonly used activity recognition datasets (MSR-Action3D [14], UTKincet-Action3D [35], MSRDailyActivity3D [44], UTD-MHAD [131], SBU Kinect Interaction [57], NTU RGB+D [84], PKU-MMD [132]) are selected for the comprasion of

different algorithms. Table 2 lists the detailed information of these RGB-D sensing based human activity datasets. Among them, NTU RGB+D and PKU-MMD are significantly larger than other datasets in terms of activity categories and samples, which makes them suitable for the evaluation of deep learning based methods. Apart from these commonly used datasets, readers may refer to [5, 133] to find more human activity recognition datasets.

The adopted RGB-D datasets are divided into three categories, namely, human action dataset, human interaction dataset, and online human activity dataset, based on the recorded action types. If the dataset contains different actions performed continually in a video stream, it is judged as an online action dataset, otherwise, if the dataset contains human interactions, it is referred as a human interaction dataset. MSR-Action3D [14] contains 20 single person actions collected in a fixed of view with a clean background. UTKincet-Action3D [35] and MSRDailyActivity3D [44] are collected for the human-object interaction purpose. By simultaneously using a Kinect sensor and an inertial

Table 2: RGB-D sensor based human activity datasets. Notation for activity types: HHI: human-human interaction, HOI: human-object interaction, SPA: single person action.

Dataset	Interactions	Subjects	Samples	Data types	Views
MSR-Action3D (2010)	20 SPA	10	567	depth(640x480) skeleton(20 joints)	1
UTKinect-Action3D (2012)	10 SPA	10	200	RGB(640x480) depth(320x240) skeleton(20 joints)	varied
MSRDailyActivity3D (2012)	16 HOI	10	320	RGB(640x480) depth(640x480) skeleton(20 joints)	1
UTD-MHAD (2015)	27 SPA	8	861	RGB(640x480) depth(320x240) skeleton(20 joints) inertial sensor signals	1
SBU Kinect Interaction (2012)	8 HHI	7	300	RGB(640x480) depth(640x480) skeleton(15 joints)	1
NTU RGB+D (2016)	11 HHI 40 HOI 9 SPA	40	56880	RGB(1920x1080) depth(512x424) skeleton (25 joints) IR sequence	3
PKU-MMD (2017)	10 HHI 41 SPA	66	21545 (1076 continuous videos)	RGB(1920x1080) depth(512x424) skeleton(25 joints) infrared sequences RGB videos	3

Table 3: Recognition performance of the state-of-the-art methods on commonly used RGB-D based human action datasets. Notation: Ref.: Reference; PDF: probability density function; RF: Random Forest; Acc.: Recognition accuracy (%).

	MSR Action3D-following evaluation protocol [14] http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/											
	Depth-based				**Skeleton-based**				**Hybrid feature-based**			
	Ref.	Year	Classifier	Acc.	Ref.	Year	Classifier	Acc.	Ref.	Year	Classifier	Acc.
Hand-crafted	[25]	(2013)	SVM	89.30	[41]	(2014)	SVM	92.46	[42]	(2014)	SVM	88.20
	[33]	(2015)	kNN	92.10	[37]	(2014)	kNN	93.61	[50]	(2018)	SVM	90.8
	[20]	(2014)	SVM	93.90	[122]	(2016)	SVM	93.96	[49]	(2017)	HMM	93.30
	[26]	(2016)	SVM	94.28	[123]	(2016)	Matching	94.40	[124]	(2016)	SVM	93.99
	[16]	(2017)	ELM	96.70	[125]	(2016)	SVM	94.4	[48]	(2013)	RF	94.30
	[27]	(2018)	SVM	**97.64**	[34]	(2018)	SVM	95.24	[30]	(2013)	SVM	94.84
					[105]	(2018)	SVM	**95.60**	[45]	(2016)	SVM	**98.20**
					[32]	(2017)	SVM	95.90				
					[126]	(2016)	Graph	**96.10**				
Deep learning	[106]	(2015)	CNN	94.58	[80]	(2015)	RNN	92.03	[104]	(2016)	CNN	84.07
	[73]	(2016)	CNN	**100.0**	[81]	(2015)	RNN	94.49	[101]	2018	CNN	94.51
					[127]	(2018)	CNN+LSTM	96.00	[103]	(2017)	RNN	**94.90**
					[128]	(2017)	LSTM	**97.22**				
	UTKinect-Action3D-following evaluation protocol [48] http://cvrc.ece.utexas.edu/KinectDatasets/HOJ3D.html											
	Depth-based				**Skeleton-based**				**Hybrid feature-based**			
	Ref.	Year	Classifier	Acc.	Ref.	Year	Classifier	Acc.	Ref.	Year	Classifier	Acc.
Hand-crafted	[27]	(2018)	SVM	86.00	[37]	(2014)	kNN	90.95	[43]	(2016)	HMM	87.90
	[21]	(2014)	PDF	**95.25**	[123]	(2016)	matching	93.47	[48]	(2013)	RF	91.90
					[126]	(2016)	Graph	95.96	[52]	(2015)	HCRF	92.00
					[41]	(2014)	SVM	97.08	[54]	(2016)	SVM	**93.90**
					[34]	(2018)	SVM	97.85				
					[122]	(2016)	SVM	**98.20**				
Deep learning	[104]	(2016)	CNN	82.00	[83]	(2017)	LSTM	95.96	[104]	(2016)	CNN	**96.00**
	[73]	(2016)	CNN	90.91	[128]	(2017)	LSTM	96.97				
	[106]	(2015)	CNN	**91.92**	[89]	(2016)	LSTM	97.00				
					[127]	(2018)	CNN+LSTM	**99.00**				
					[88]	(2018)	LSTM	99.00				
	MSRDailyActivity3D-following evaluation protocol [44] http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/											
	Depth-based				**Skeleton-based**				**Hybrid feature-based**			
	Ref.	Year	Classifier	Acc.	Ref.	Year	Classifier	Acc.	Ref.	Year	Classifier	Acc.
Hand-crafted	[19]	(2013)	SVM	80.00	[31]	(2013)	kNN	73.80	[124]	(2016)	SVM	73.21
	[20]	(2014)	SVM	86.25	[32]	(2017)	SVM	75.00	[50]	(2018)	SVM	81.30
	[22]	(2016)	SVM	80.63	[129]	(2016)	MIL	78.52	[54]	(2016)	SVM	86.00
	[16]	(2017)	ELM	**89.00**	[27]	(2018)	SVM	**91.00**	[56]	(2016)	DRRL	87.50
									[45]	(2016)	SVM	91.25
									[47]	(2014)	SVM	93.10
									[49]	(2017)	HMM	**94.10**
Deep learning	[106]	(2015)	CNN	78.12	[127]	(2018)	CNN+LSTM	**63.10**				
	[73]	(2016)	CNN	85.00								
	[130]	(2017)	CNN+LSTM	**86.90**								

sensor to capture human actions, UTD-MHAD [131] dataset explores the possibility in fusing different sources of data to improve the recognition performance. SBU Kinect Interaction [57] is recorded for the study of human-human interactions in a laboratory environment. In NTU RGB+D [84], a large number of single human actions, human-object interactions, and human-human interactions are collected. PKU-MMD [132] provides over a thousand videos involving continuous actions for online human activity understanding.

Table 3 categorizes techniques and compares their performance on three commonly used human action datasets to help select suitable techniques for particular applications. Each column of the table contains one type of methods, i.e., depth-based, skeleton-based or hybrid feature-based methods. Inside the column, the algorithms are further ranked according to the achieved accuracy. It can be seen that all the three categories of methods have achieved good recognition performance on the MSR Action 3D dataset due to its simplified experimental setting and action classes. Among them, 100% accuracy is obtained by [73] which converted the classic DMM to RGB images and utilized CNN for classification. However, their performance decreases greatly in different viewpoint settings due to the dramatic variation of depth maps. This depth image's viewing angle sensitivity problem can be also observed by comparing the first and second column of the UTKincet-Action3D dataset collected in three views, where most of the skeleton-based methods achieve overwhelming accuracy than the depth-based methods. Actually, based on the accuracy on this dataset, it is also easy to find that the skeleton-based methods are better suited for the classification of actions under different viewing angles than the depth-based methods and hybrid features-based methods. On the other hand, the hybrid features-based approaches outperform the skeleton-based or depth-based methods in the human-object interaction dataset of MSRDailyActivity3D, indicating that the skeleton alone is insufficient to distinguish actions which involve human-object interactions. The reason might be that the contexture information of objects also plays an important role in the defined actions.

Table 3 also divides the methods into hand-crafted methods and deep learning methods. The table shows that the top recognition accuracy of MSR Action3D dataset and UTKinect-Action3D dataset are both achieved by deep learning based methods,

Table 4: Comparison of state of the art methods on the **UTD-MHAD** dataset in terms of accuracies, classifier types, and sensor types. Notation: Acc.: Accuracy(%); K: Kinect sensor; I: Inertial sensor; CRC: Collaborative Representation Classifier; MBC: Multi-Class Boosting; MVS: Multi-View Stacking.

UTD-MHAD (cross-subject [131]) http://www.utdallas.edu/~kehtar/UTD-MHAD.html									
Hand-crafted					Deep learning				
Ref.	Year	Sensor	Classifier	Acc.	Ref.	Year	Sensor	Classifier	Acc.
[131]	(2015)	K	CRC	66.10	[74]	(2016)	K	CNN	86.97
[131]	(2015)	K+I	CRC	79.10	[70]	(2017)	K	CNN	88.10
[17]	(2017)	K	MBC	84.40	[101]	(2018)	K	CNN	88.14
[134]	(2018)	K	SVM	92.00	[79]	(2018)	K	CNN	94.51
[134]	(2018)	K+I	SVM	96.10	[135]	(2017)	K	CNN	96.27
[136]	(2018)	K	MVS	90.90	[72]	(2017)	K	CNN	97.20
[136]	(2018)	K+I	MVS	**98.10**	[137]	(2018)	K	CNN	**97.90**

which demonstrates their effectiveness in human action recognition. On the other hand, it can also be observed that the highest performance of hand crafted-based methods has also reached 98.2% accuracy ([45], [122]) on both datasets, leaving few spaces for further development. Compared to the former two datasets, fewer deep learning based methods are evaluated on the MSRDaliyAcitivity3D dataset and hand-crafted methods achieve better performance at this stage. Regarding the classifier, most of the hand-crafted methods adopt the SVM, while deep learning methods normally use CNN, LSTM or their combination for recognition.

Table 4 shows a comparison of human action recognition performance on the multi-model UTD-MHAD dataset to analyze the feasibility in combining data from different sensors to boost the recognition performance. It can be easily observed that the benefit of combining the Kinect sensor and inertial sensor is overwhelming among the hand-crafted based methods. For example, both [131] and [136] achieved over 7% of accuracy improvement by combining the data from the Kinect sensor and inertial sensor. Using an effective fusing strategy, the hand-crafted based methods have achieved similar performance with the deep learning based methods (98.1% and 97.9%, respectively). Its should also be noted that most of the existing deep learning based methods only used the data from the Kinect sensor. Thus, their performance might be improved by jointly using the data from the both sensors.

Table 5 reports a comparison of the state-of-the-art methods on two commonly

Table 5: Comparison of the state-of-the-art methods on the **SBU Kinect Interaction** and **NTU RGB+D** dataset in terms of accuracies, method types, and solutions to different challenges. Notation: Acc.: Accuracy(%); S: skeleton; Separately: consider each person's action as an individual sample and averaging the classification scores for the final prediction; Maxout: use an element-wise maximum operation to merge two persons' feature maps in the designed network structure; Concatenation: simply stack each person's feature together or further include the interrelationship between the two persons.

SBU Kinect Interaction(5-fold cross-validation [57]) (http://www3.cs.stonybrook.edu/~kyun/research/kinect_interaction/)				
Ref.	Year	Acc.	Type	Interaction solution
[57]	2012	80.30	(S+MIL)	Distance
[59]	2014	86.90	S+SVM	Body part
[60]	2015	89.40	(S+SVM)	Body part
[61]	2017	91.00	(S+SLM)	Distance
[138]	2017	**91.12**	(S+SVM)	Body part
[86]	2016	90.4	S+LSTM	Seperately
[90]	2017	91.5	S+LSTM	Concatenation
[89]	2016	93.3	S+LSTM	Concatenation
[71]	2017	93.6	S+CNN	Seperately
[91]	2017	94.8	S+RNN	Seperately
[88]	2018	94.90	S+LSTM	Concatenation
[87]	2018	97.2	S+RNN	Concatenation
[83]	2017	**99.0**	S+LSTM	Concatenation

NTU RGB+D(following evaluation protocol [84]) (http://rose1.ntu.edu.sg/datasets/actionrecognition.asp)					
Ref.	Year	Acc.		Type	Interaction solution
		cross-subject	cross-view		
[84]	2016	62.93	70.27	S+LSTM	Concatenation
[127]	2018	67.50	76.21	S+CNN+LSTM	Concatenation
[77]	2017	69.20	77.70	S+CNN	Concatenation
[83]	2017	70.26	82.39	S+LSTM	Concatenation
[91]	2017	71.30	79.50	S+RNN	Seperately
[90]	2017	73.40	81.20	S+LSTM	Concatenation
[128]	2017	74.60	81.25	S+LSTM	Concatenation
[97]	2017	75.20	83.10	S+depth	Concatenation
[88]	2018	76.10	84.00	S+LSTM	Concatenation
[70]	2017	76.20	82.30	S+CNN	Concatenation
[87]	2017	79.40	87.60	S+RNN	Concatenation
[139]	2018	79.5	87.6	S+RNN	Concatenation
[71]	2017	79.56	84.83	S+CNN	Seperately
[140]	2017	80.03	87.21	S+CNN	Concatenation
[72]	2017	81.30	89.20	S+CNN	Concatenation
[76]	2018	81.50	88.30	S+GCN	Concatenation
[141]	2018	83.5	89.8	S+GCN	Maxout
[75]	2018	82.67	93.22	S+RNN+CNN	Concatenation
[92]	2018	84.80	92.4	S+RNN+CNN	Concatenation
[79]	2018	**91.71**	**95.26**	S+CNN	Concatenation

used human interaction datasets: SBU Kinect Interaction and NTU RGB+D, in terms of accuracies, method types, and solutions to the interaction challenge. The interaction challenge lies in the adapting of single human's action features into a representation that is suitable for the human interaction scenario. On the SBU Kinect Interaction dataset, the top performance of deep learning based methods (99.0%, [83]) outperforms the top hand-crafted based method (91.12%, [138]) to a large extent, indicating that the deep learning based methods are better suited for human interaction recognition. It can also be observed that most of the existing methods on the NTU RGB+D dataset are based on deep learning technologies.

Table 5 also shows that most of the human interaction approaches are based on the skeleton data rather than depth images. In hand-crafted methods, the inter-relationship between two persons is modeled by using interactive body parts or joints distance information. While in the deep learning based methods, this challenge is handled by a concatenation operation, maxout operation or the simple separation strategy which recognizes each person's action and averages the classification scores for the final prediction. In the concatenation operation, similar technologies in hand-crafted methods can be explored to further improve the recognition performance. A prominent example is [83], which modeled the skeleton and the bone relationship between two people before inputting into a three layer LSTM network and achieved 99% accuracy on the SBU Kinect Interaction dataset. [77] also shows that deep learning based methods can use the similar feature (the lie group) with the hand-crafted method [41].

Table 6: Online performance of the state-of-the-art methods on the **PKU-MMD dataset**. Notation: mAP (%): mean Average Precision under a threshold θ; S: Skeleton.

PKU-MMD following the evaluation protocol [132] http://www.icst.pku.edu.cn/struct/Projects/PKUMMD.html							
Methods	Year	Type	Cross-subject (mAP)		Cross-view (mAP)		Detection operation
			$\theta = 0.1$	$\theta = 0.5$	$\theta = 0.1$	$\theta = 0.5$	
[120]	2016	S+RNN	45.2	32.5	69.9	53.3	sliding window
[132]	2017	S+LSTM	47.9	13.0	54.5	15.9	sliding window
[121]	2018	S+LSTM	51.3	35.2	63.2	48.6	action proposal
[139]	2018	S+RNN	**87.4**	81.1	**95.3**	91.1	sliding window
[142]	2018	S+CNN	-	**92.6**	-	**94.2**	action proposal

To make the action recognition problem simpler, each action video in the previous

six datasets is manually trimmed to contain only one complete activity. However, in many practical scenarios, it is hard to know the exact starting time and ending time of an action ahead. The algorithms are required to continuously output the recognition results even when the activity is still ongoing. The big gap between the simplified scenarios and practical scenarios makes the online performance of the existing methods unclear when applied in real-world applications. PKU-MMD was collected to provide continuous activity videos for the study of online activity recognition. Table 6 compares the existing methods on the PKU-MMD dataset in terms of detection accuracies and detection strategies. The accuracy is measured by mean average precision (mAP) [143] which evaluates the detection precision of different overlapping ration between the predicted interval and the ground truth interval. As shown in Table 6, existing online action recognition methods usually use a sliding window strategy or an action proposal strategy for action detection. The action proposals are normally generated by training an extra network [121]. Compared to the sliding window strategy which might yield noisy predictions for some frames, the action proposal based solutions can achieve a more stable detection performance.

It can be observed from the tables that deep learning based methods have achieved overwhelming recognition performance over hand-crafted based methods in most of the existing human activity datasets. However, it is also well-known that most of the deep learning based approaches require large training samples to reduce the affect of overfitting and achieve better performance. Existing solutions to this problem mainly focus on three aspects: 1) finetune the models pre-trained on larger datasets [70, 79]; 2) randomly crop sub-sequences from an entire sequence [86]; 3) adopt synthetic data with existing data to improve the performance. For example, Rahmani *et al.* [93] proposed to learn a view-invariant pose model with the depth images synthesized from a small number of human poses. Thus, enhancing existing datasets to boost the performance of CNNs remains a great potential future direction.

Apart from the recognition accuracy, it is also essential for the algorithms to be computationally efficient for many real-world motion recognition applications. Most of the existing hand-crafted methods achieved real-time performance via a careful design of the features and the use of low computational cost classifiers such as SVM [27, 122,

126]. Due to the complex structure of neural networks, existing deep learning based methods heavily rely on advanced parallel computing devices such as GPU and TPU to reach real time performance. Thus, exploring effective light weighted networks is a good solution to relieve the computational burden.

7. Conclusion and Future Directions

In this paper, we have provided a comprehensive analysis of RGB-D sensing based human action and interaction recognition, ranging from hand-crafted algorithms to deep learning algorithms. While significant progress has been achieved in improving the recognition accuracy, there remains great challenges such as online adaption, viewpoint variations, occlusions, and action duration variations. Along with existing solutions, these challenges have also been investigated in detail in this paper. The future directions are summarized as follows:

Fusion of multi-modal data. Multi-modal data is beneficial for human activity recognition, mainly because it not only provides richer information but also it can be used for reducing the noises in single source data and improving the robustness of the recognition performance. Thus, for the future research on human activities, more effective integration of diverse information should be developed instead of the monotonous concatenation of features from different sources. For human interactions, the fusion of features from individuals and correlations extracted from various data sources might produce the more robust interpretation. In addition, the contextual information from the surrounding environment which is relatively unexplored could enhance the performance of traditional feature representation for human action recognition.

Development of view-invariance algorithms. Tolerance to different viewing angles is an useful property since it not only allows the subjects to move around but it also removes extra calibration procedures for different sensor locations. The skeleton-based methods have an inherent resistance towards different viewing angles, however, the estimated skeleton data might not be accurate in side views, which will probably result in a drop of recognition performance. Current efforts of depth-based methods in this direction mainly focus on generating synthetic multi-view data to augment the training

samples. Thus, future research may devote more attention to develop view-invariant feature descriptors.

Evaluation on practical scenarios. Evaluation of the activity recognition algorithms on practical scenarios is yet only a partially solved problem since most of the existing RGB-D datasets are collected in constrained environments. There is a big gap between the collected datasets and the wild environment due to the insufficient categories, samples and occlusion cases, restricted actions, limited distance variations and constrained indoor environment settings. This makes it hard for algorithms to be generalized to practical situations in the real world. Therefore, the collection of large-scale action datasets for both training and evaluation for practical scenarios should be one future direction.

Learning directly from raw video data. Although many deep learning methods could outperform hand-crafted methods, most of them require a pre-processing step to extract hand-crafted representations from RGB, depth or skeleton data. While the handcrafted representations simply the feature dimensions to a large extent, they also limit the interpretation ability of deep learning methods. Currently, this compromise might be due to the insufficient training samples. Thus, given sufficient training samples, another future direction will be to develop novel deep learning architectures to learn representations directly from raw video data.

Online activity recognition. While large attentions have been focused on developing highly accurate activity recognition algorithms for pre-segmented video sequences, the online recognition system, which aims to analyze human behaviors instantly from a continuous video stream, is indeed demanded by practical applications. Moreover, since most of the current research are evaluated on trimmed data where each segment contains one whole category, it still remains unclear about their performance when applied to online cases. Therefore, developing recognition approaches that can be applied to practical scenarios is an essential direction.

Interpretation of human behavior. Human behavior, which consists of many components such as human activity, facial expressions, visual focus of attention etc, is more complicated than human action or interaction. Automatic interpretation of human behavior is an essential step towards developing real intelligent systems and is

beneficial in the application that explores the human cognitive status.

Acknowledgments

This work was supported in part by the EU Seventh Framework Programme (No. 611391, Development of Robot-Enhanced therapy for children with AutisM spectrum disorders (DREAM)) and China Scholarship Council.

References

[1] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: a literature review, Pattern Recog. 8 (48) (2015) 2329–2345.

[2] J. K. Aggarwal, M. S. Ryoo, Human activity analysis: A review, ACM Computing Surveys (CSUR) 43 (3) (2011) 16.

[3] E. Cippitelli, F. Fioranelli, E. Gambi, S. Spinsante, Radar and rgb-depth sensors for fall detection: a review, IEEE Sensors Journal 17 (12) (2017) 3585–3604.

[4] Y. Kong, Y. Fu, Modeling supporting regions for close human interaction recognition, in: ECCV, 2014, pp. 29–44.

[5] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang, Rgb-d-based action recognition datasets: A survey, Pattern Recog. 60 (2016) 86–105.

[6] L. L. Presti, M. La Cascia, 3d skeleton-based human action classification: A survey, Pattern Recognition 53 (2016) 130–147.

[7] J. Aggarwal, L. Xia, Human activity recognition from 3d data: A review, Pattern Recog. Lett. 48 (2014) 70–80.

[8] L. Chen, H. Wei, J. Ferryman, A survey of human motion analysis using depth imagery, Pattern Recog. Lett. 34 (15) (2013) 1995–2006.

[9] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, Time-of-Flight and Depth Imaging. Sensors, Algorithms, Appl. (2013) 149–187.

[10] F. Han, B. Reily, W. Hoff, H. Zhang, Space-time representation of people based on 3d skeletal data: A review, Comput. Vis. Image Understanding 158 (2017) 85–105.

[11] F. Zhu, L. Shao, J. Xie, Y. Fang, From handcrafted to learned representations for human action recognition: A survey, Image and Vis. Comput. 55 (2016) 42–52.

[12] R. Lun, W. Zhao, A survey of applications and human motion recognition with microsoft kinect, Int. J. Pattern Recog. Artificial Intell. 29 (05) (2015) 1555008.

[13] C. Chen, K. Liu, N. Kehtarnavaz, Real-time human action recognition based on depth motion maps, J. Real-Time Image Processing (2013) 1–9.

[14] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: CVPRW, 2010, pp. 9–14.

[15] M. F. Bulbul, Y. Jiang, J. Ma, Human action recognition based on dmms, hogs and contourlet transform, in: Proc. IEEE Int. Conf. Multimedia Big Data, 2015, pp. 389–394.

[16] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, N. Kehtarnavaz, Multi-temporal depth motion maps-based local binary patterns for 3-d human action recognition, IEEE Access 5 (2017) 22590–22604.

[17] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, L. Shao, Action recognition using 3d histograms of texture and a multi-class boosting classifier, IEEE Trans. Image Process. 26 (10) (2017) 4648–4660.

[18] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proc. ACM Int. Conf. Multimedia, 2012, pp. 1057–1060.

[19] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: CVPR, 2013, pp. 716–723.

[20] X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in: CVPR, 2014, pp. 804–811.

[21] R. Slama, H. Wannous, M. Daoudi, Grassmannian representation of motion depth for 3d human gesture and action recognition, in: ICPR, 2014, pp. 3499–3504.

[22] C. Jia, Y. Fu, Low-rank tensor subspace learning for rgb-d action recognition, IEEE Trans. Image Process. 25 (10) (2016) 4641–4652.

[23] C. Jia, Y. Kong, Z. Ding, Y. R. Fu, Latent tensor transfer learning for rgb-d action recognition, in: Proc. 22rd ACM Int. Conf. Multimedia, 2014, pp. 87–96.

[24] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3d action recognition with random occupancy patterns, in: ECCV, 2012, pp. 872–885.

[25] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: CVPR, 2013, pp. 2834–2841.

[26] M. Liu, H. Liu, Depth context: a new descriptor for human activity recognition by using sole depth sequences, Neurocomputing 175 (2016) 747–758.

[27] M. Liu, H. Liu, C. Chen, Robust 3d action recognition through sampling local appearances and global distributions, IEEE Trans. Multimedia 20 (8) (2018) 1932–1947.

[28] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.

[29] M. A. Gowayyed, M. Torki, M. E. Hussein, M. El-Saban, Histogram of oriented displacements (hod): describing trajectories of human joints for action recognition, in: Proc. Int. joint Conf. Artificial Intell., 2013, pp. 1351–1357.

[30] E. Ohn-Bar, M. M. Trivedi, Joint angles similarities and hog2 for action recognition, in: CVPRW, 2013, pp. 465–470.

[31] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection, in: ICCV, 2013, pp. 2752–2759.

[32] R. Qiao, L. Liu, C. Shen, A. van den Hengel, Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recog. 66 (2017) 202–212.

[33] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A. Del Bimbo, 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold, IEEE Trans. Cybern. 45 (7) (2015) 1340–1352.

[34] Y. Guo, Y. Li, Z. Shao, Dsrf: A flexible trajectory descriptor for articulated human action recognition, Pattern Recog. 76 (2018) 137–148.

[35] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: CVPRW, 2012, pp. 20–27.

[36] H. Pazhoumand-Dar, C.-P. Lam, M. Masek, Joint movement similarities for robust 3d action recognition using skeletal data, J. Vis. Commun. Image Represent. 30 (2015) 10–21.

[37] I. Theodorakopoulos, D. Kastaniotis, G. Economou, S. Fotopoulos, Pose-based human action recognition via sparse representation in dissimilarity space, J. Vis. Commun. Image Represent. 25 (1) (2014) 12–23.

[38] I. Lillo, J. C. Niebles, A. Soto, Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos, Image Vis. Comput. 59 (2017) 63–75.

[39] A. Eweiwi, M. S. Cheema, C. Bauckhage, J. Gall, Efficient pose-based action recognition, in: Asian Conf. Comput. Vis., 2015, pp. 428–443.

[40] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, F. Flórez-Revuelta, Evolutionary joint selection to improve human action recognition with rgb-d devices, Expert systems with applications Expert Syst. Appl. 41 (3) (2014) 786–794.

[41] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: CVPR, 2014, pp. 588–595.

[42] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3d human action recognition, IEEE Trans. Pattern Anal. Mach. Intell 36 (5) (2014) 914–927.

[43] N. Raman, S. Maybank, Activity recognition using a supervised non-parametric hierarchical hmm, Neurocomputing 199 (2016) 163–177.

[44] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: CVPR, 2012, pp. 1290–1297.

[45] A. Shahroudy, T.-T. Ng, Q. Yang, G. Wang, Multimodal multipart learning for action recognition in depth videos, IEEE Trans. Pattern Anal. and Mach. Intell. 38 (10) (2016) 2123–2129.

[46] X. Yang, Y. Tian, Effective 3d action recognition using eigenjoints, J. Vis. Commun. Image Represent. 25 (1) (2014) 2–11.

[47] S. Althloothi, M. H. Mahoor, X. Zhang, R. M. Voyles, Human activity recognition using multi-features and multiple kernel learning, Pattern Recog. 47 (5) (2014) 1800–1812.

[48] Y. Zhu, W. Chen, G. Guo, Fusing spatiotemporal features and joints for 3d action recognition, in: CVPRW, 2013, pp. 486–491.

[49] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, D. Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, Pattern Recog. 61 (2017) 295–308.

[50] X. Ji, J. Cheng, W. Feng, D. Tao, Skeleton embedded motion body partition for human action recognition using depth sequences, Signal Processing 143 (2018) 56–68.

[51] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from rgbd images, in: ICRA, 2012, pp. 842–849.

[52] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, Z.-X. Yang, Coupled hidden conditional random fields for rgb-d human action recognition, Signal Processing 112 (2015) 74–82.

[53] Y. Kong, Y. Fu, Bilinear heterogeneous information machine for rgb-d action recognition, in: CVPR, 2015, pp. 1054–1062.

[54] H. Zhang, L. E. Parker, Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos, IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2016) 541–555.

[55] Y. Kong, Y. Fu, Max-margin heterogeneous information machine for rgb-d action recognition, Inter. J. Comput. Vis. 123 (3) (2017) 350–371.

[56] Y. Kong, Y. Fu, Discriminative relational representation learning for rgb-d action recognition, IEEE Trans. Image Process. 25 (6) (2016) 2856–2865.

[57] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: CVPRW, 2012, pp. 28–35.

[58] T. Huynh-The, O. Banos, B.-V. Le, D.-M. Bui, S. Lee, Y. Yoon, T. Le-Tien, Pam-based flexible generative topic model for 3d interactive activity recognition, in: IEEE Int. Conf. Advanced Technol. Commun., 2015, pp. 117–122.

[59] Y. Ji, G. Ye, H. Cheng, Interactive body part contrast mining for human interaction recognition, in: IEEE Int. Conf. Multimedia and Expo Workshops, 2014, pp. 1–6.

[60] Y. Ji, H. Cheng, Y. Zheng, H. Li, Learning contrastive feature distribution model for interaction recognition, J. Vis. Commun. Image Represent. 33 (2015) 340–349.

[61] H. Wu, J. Shao, X. Xu, Y. Ji, F. Shen, H. T. Shen, Recognition and detection of two-person interactive actions using automatically selected skeleton features, IEEE Trans. Human Mach. Syst. 48 (3) (2018) 304–310.

[62] V. Bloom, V. Argyriou, D. Makris, Hierarchical transfer learning for online recognition of compound actions, Comput. Vis. Image Understanding 144 (2016) 62–72.

[63] T. Hu, X. Zhu, W. Guo, K. Su, Efficient interaction recognition through positive action representation, Math. Problems in Eng. 2013 (2013) 1–11.

[64] C. Coppola, D. R. Faria, U. Nunes, N. Bellotto, Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data, in: IROS, 2016, pp. 5055–5061.

[65] I. Gori, J. Aggarwal, L. Matthies, M. S. Ryoo, Multi-type activity recognition from a robot's viewpoint, in: Proc. Int. Joint Conf. Artificial Intell., 2017, pp. 4849–4853.

[66] C. van Gemeren, R. T. Tan, R. Poppe, R. C. Veltkamp, Dyadic interaction detection from pose and flow, Human Behavior Understanding (2014) 101–115.

[67] L. Xia, I. Gori, J. Aggarwal, M. Ryoo, Robot-centric activity recognition from first-person rgb-d videos, in: IEEE Winter Conf. Applicat. Comput. Vis., 2015, pp. 357–364.

[68] R. Alazrai, Y. Mowafi, C. G. Lee, Anatomical-plane-based representation for human–human interactions analysis, Pattern Recog. 48 (8) (2015) 2346–2363.

[69] R. Trabelsi, J. Varadarajan, Y. Pei, L. Zhang, I. Jabri, A. Bouallegue, P. Moulin, Robust multi-modal cues for dyadic human interaction recognition, in: Proc. Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, 2017, pp. 47–53.

[70] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Processing Lett. 24 (5) (2017) 624–628.

[71] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: CVPR, 2017, pp. 4570–4579.

[72] J. Liu, N. Akhtar, A. Mian, Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition, arXiv preprint arXiv:1711.05941, 2017.

[73] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P. O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Human Mach. Syst. 46 (4) (2016) 498–509.

[74] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra based action recognition using convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. 28 (3) (2018) 807–811.

[75] C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, J. Liu, Memory attention networks for skeleton-based action recognition, arXiv preprint arXiv:1804.08254, 2018.

[76] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, 2018.

[77] Z. Huang, C. Wan, T. Probst, L. Van Gool, Deep learning on lie groups for skeleton-based action recognition, in: CVPR, 2017, pp. 6099–6108.

[78] R. Vemulapalli, R. Chellapa, Rolling rotations for recognizing human actions from 3d skeletal data, in: CVPR, 2016, pp. 4471–4479.

[79] M. Liu, J. Yuan, Recognizing human actions as the evolution of pose estimation maps, in: CVPR, 2018, pp. 1159–1168.

[80] V. Veeriah, N. Zhuang, G.-J. Qi, Differential recurrent neural networks for action recognition, in: ICCV, 2015, pp. 4041–4049.

[81] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: CVPR, 2015, pp. 1110–1118.

[82] Y. Du, Y. Fu, L. Wang, Representation learning of temporal dynamics for skeleton-based action recognition, IEEE Trans. Image Process. 25 (7) (2016) 3010–3022.

[83] S. Zhang, X. Liu, J. Xiao, On geometric features for skeleton-based action recognition using multilayer lstm networks, in: WACV, 2017, pp. 148–157.

[84] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: CVPR, 2016, pp. 1010–1019.

[85] B. Mahasseni, S. Todorovic, Regularizing long short term memory with 3d human-skeleton sequences for action recognition, in: CVPR, 2016, pp. 3054–3062.

[86] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks, in: AAAI, 2016, pp. 3697–3703.

[87] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data, in: ICCV, 2017, pp. 2117–2126.

[88] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, A. C. Kot, Skeleton-based human action recognition with global context-aware attention lstm networks, IEEE Trans. Image Process. 27 (4) (2018) 1586–1599.

[89] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: ECCV, 2016, pp. 816–833.

[90] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data., in: AAAI, 2017, pp. 4263–4270.

[91] H. Wang, L. Wang, Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks, CVPR (2017) 499–508.

[92] C. Si, Y. Jing, W. Wang, L. Wang, T. Tan, Skeleton-based action recognition with spatial reasoning and temporal stack learning, in: ECCV, 2018, pp. 103–118.

[93] H. Rahmani, A. Mian, 3d action recognition from novel viewpoints, in: CVPR, 2016, pp. 1506–1515.

[94] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, X. Cao, Z. Liu, X. Chai, Z. Liu, et al., Multimodal gesture recognition based on the resc3d network, in: CVPR, 2017, pp. 3047–3055.

[95] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, M. Bennamoun, Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition, in: CVPR, 2017, pp. 3120–3128.

[96] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, J.-M. Odobez, Deep dynamic neural networks for multimodal gesture segmentation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1583–1597.

[97] H. Rahmani, M. Bennamoun, Learning action recognition model from depth and skeleton videos, in: CVPR, 2017, pp. 5832–5841.

[98] E. P. Ijjina, K. M. Chalavadi, Human action recognition in rgb-d videos using motion sequence information and deep learning, Pattern Recog. 72 (2017) 504–516.

[99] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell 35 (1) (2013) 221–231.

[100] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: ICCV, 2015, pp. 4489–4497.

[101] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, D. D. Feng, Deep convolutional neural networks for human action recognition using depth maps and postures, IEEE Trans. Syst. Man, Cybern.: Syst. (99) (2018) 1–14.

[102] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, P. Ogunbona, Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks, in: CVPR, 2017, pp. 595–604.

[103] Z. Shi, T.-K. Kim, Learning and refining of privileged information-based rnns for action recognition from depth sequences, in: CVPR, 2017, pp. 3461–3470.

[104] Z. Liu, C. Zhang, Y. Tian, 3d-based deep convolutional neural network for action recognition with depth sequences, Image Vis. Comput. 55 (2016) 93–100.

[105] B. Liu, Z. Ju, H. Liu, A structured multi-feature representation for recognizing human action and interaction, Neurocomputing 318 (2018) 287–296.

[106] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P. Ogunbona, Deep convolutional neural networks for action recognition using depth map sequences, arXiv preprint arXiv:1501.04686, 2015.

[107] A. Chaaraoui, J. Padilla-Lopez, F. Flórez-Revuelta, Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices, in: ICCV Workshops, 2013, pp. 91–97.

[108] J.-W. Hsieh, S.-Y. Chen, C.-H. Chuang, M.-F. Chueh, S.-S. Yu, Occluded human body segmentation and its application to behavior analysis, in: Proc. IEEE Int. Symp. Circuit and Syst., 2010, pp. 3433–3436.

[109] N.-G. Cho, A. L. Yuille, S.-W. Lee, Adaptive occlusion state estimation for human pose tracking under self-occlusions, Pattern Recog. 46 (3) (2013) 649–661.

[110] Y. Kong, Y. Fu, Close human interaction recognition using patch-aware models, IEEE Trans. Image Process. 25 (1) (2016) 167–178.

[111] H. S. Koppula, A. Saxena, Anticipating human activities using object affordances for reactive robotic response, IEEE Trans. Pattern Anal. Mach. Intell 38 (1) (2016) 14–29.

[112] P. Banerjee, R. Nevatia, Pose filter based hidden-crf models for activity detection, in: ECCV, 2014, pp. 711–726.

[113] G. Zhu, L. Zhang, P. Shen, J. Song, An online continuous human action recognition algorithm based on the kinect sensor, Sensors 16 (2) (2016) 161.

[114] D. Huang, S. Yao, Y. Wang, F. De La Torre, Sequential max-margin event detectors, in: ECCV, 2014, pp. 410–424.

[115] C. Wu, J. Zhang, S. Savarese, A. Saxena, Watch-n-patch: Unsupervised understanding of actions and relations, in: CVPR, 2015, pp. 4362–4370.

[116] M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, A. Del Bimbo, Motion segment decomposition of rgb-d sequences for human behavior understanding, Pattern Recog. 61 (2017) 222–233.

[117] D. Gong, G. Medioni, X. Zhao, Structured time series analysis for human action segmentation and recognition, IEEE Trans. Pattern Anal. Mach. Intell 36 (7) (2014) 1414–1427.

[118] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network, in: CVPR, 2016, pp. 4207–4215.

[119] Z. Shou, D. Wang, S. Chang, Action temporal localization in untrimmed videos via multi-stage cnns, in: CVPR, Vol. 3, 2016.

[120] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, J. Liu, Online human action detection using joint classification-regression recurrent neural networks, in: ECCV, 2016, pp. 203–220.

[121] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Spatio-temporal attention based lstm networks for 3d action recognition and detection, IEEE Trans. Image Process. 27 (7) (2018) 3459–3471.

[122] P. Koniusz, A. Cherian, F. Porikli, Tensor representations via kernel linearization for action recognition from 3d skeletons, in: ECCV, 2016, pp. 37–53.

[123] C. Wang, Y. Wang, A. L. Yuille, Mining 3d key-pose-motifs for action recognition, in: CVPR, 2016, pp. 2639–2647.

[124] Y. Kong, B. Satarboroujeni, Y. Fu, Learning hierarchical 3d kernel descriptors for rgb-d action recognition, Comput. Vis. Image Understanding 144 (C) (2016) 14–23.

[125] B. Liu, H. Yu, X. Zhou, D. Tang, H. Liu, Combining 3d joints moving trend and geometry property for human action recognition, in: IEEE Int. Conf. Syst. Man, Cyber., 2016, pp. 000332–000337.

[126] H. Chen, G. Wang, J.-H. Xue, L. He, A novel hierarchical framework for human action recognition, Pattern Recog. 55 (2016) 148–159.

[127] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, J. F. Vélez, Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition, Pattern Recog. 76 (2018) 80–94.

[128] I. Lee, D. Kim, S. Kang, S. Lee, Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks, in: ICCV, 2017, pp. 1012–1020.

[129] X. Cai, W. Zhou, L. Wu, J. Luo, H. Li, Effective active skeleton representation for low latency human action recognition, IEEE Trans. Multimedia 18 (2) (2016) 141–154.

[130] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, L. Fei-Fei, Unsupervised learning of long-term motion dynamics for videos, in: CVPR, 2017.

[131] C. Chen, R. Jafari, N. Kehtarnavaz, Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: ICIP, 2015, pp. 168–172.

[132] C. Liu, Y. Hu, Y. Li, S. Song, J. Liu, Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding, arXiv preprint arXiv:1703.07475, 2017.

[133] S. Escalera, X. Baró, H. J. Escalante, I. Guyon, Chalearn looking at people: A review of events and resources, in: Int. Joint Conf. Neural Networks, 2017, pp. 1594–1601.

[134] N. E. D. H. Elmadany, Y. He, L. Guan, Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis, IEEE Trans. Multimedia (2018) 1–14.

[135] B. Li, M. He, X. Cheng, Y. Chen, Y. Dai, Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn, in: IEEE Int. Conf. Multimedia, 2017, pp. 601–604.

[136] E. Garcia-Ceja, C. E. Galván-Tejada, R. Brena, Multi-view stacking for activity recognition with sound and accelerometer data, Information Fusion 40 (2018) 45–56.

[137] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu, Y. Zhang, Skeleton-based action recognition with gated convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. (2018) 1–11.

[138] B. Liu, H. Cai, X. Ji, H. Liu, Human-human interaction recognition based on spatial and motion trend feature, in: ICIP, 2017, pp. 4547–4551.

[139] H. Wang, L. Wang, Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection, IEEE Trans. Image Process. 27 (9) (2018) 4382–4394.

[140] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recog. 68 (2017) 346–362.

[141] Y. Tang, Y. Tian, J. Lu, P. Li, J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, in: CVPR, 2018, pp. 5323–5332.

[142] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in: IJCAI, 2018, pp. 1–8.

[143] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Inter. J. Comput. Vis. 88 (2) (2010) 303–338.